SPECIAL REPORT

Seize the E-Journal: Models for Archiving symposium: report

By Ramune K. Kubilius, MALS, AHIP
r-kubilius@northwestern.edu
Cochair
Symposium Planning Committee
MLA Collection Development Section and
Collection Development/Special Projects Librarian

Galter Health Sciences Library
Northwestern University
303 East Chicago Avenue
Chicago, Illinois 60611-3093

Linda J. Walton, MLS
ljwalton@northwestern.edu
Cochair
Symposium Planning Committee
MLA Collection Development Section and
Associate Director

Galter Health Sciences Library
Northwestern University
303 East Chicago Avenue
Chicago, Illinois 60611-3093

INTRODUCTION

A half-day symposium, ''Seize the E-Journal: Models for Archiving,'' was held, May 26, 2004, after the conclusion of MLA '04, the 104th Annual Meeting of the Medical Library Association (MLA), in Washington, DC. The symposium was sponsored by the Collection Development Section.

The objectives of the symposium were for the 100 participants to become familiar with current electronic journal archiving models, to discuss future-oriented priorities for collection development, and to suggest electronic archival strategies for individual, organizational, consortial, and national libraries. The program included presentations by four experts on their archiving models, input from reaction panel members on the various models presented, a question-and-answer session, and, finally, a breakout discussion session that allowed participants to brainstorm on future archiving priorities and standards. Presentations from the session, along with a Web bibliography of related resources, are maintained on the Collection Development Section Website <http://colldev.mlanet.org>.

PLANNING

Every few years, the Collection Development Section sponsors a symposium on a collections topic before or after an MLA annual meeting. As the section considered ''hot'' topics for a symposium, it was clear that librarians' role as preservationists of biomedical information is being challenged. In many cases, information acquired by the library is housed on a host site such as a publisher's site or an aggregator's site. New models for preserving electronic information are being developed and tested, so that biomedical information access will continue into the future. As more and more medical libraries cancel their print subscriptions, the various electronic preservation options need to be understood, so that informed decisions are made as they relate to perpetual access to licensed information. Some of the issues being considered are:

- Who is responsible for preserving this information?
- What technology is being used to manage this information?
- Who is making the decisions on what is to be preserved?

Once the symposium proposal was accepted by the MLA Continuing Education Committee, a planning committee made up of section members was formed. Research was conducted to identify current and planned models for preservation and archiving of online resources. Four models were selected to represent current trends in online preservation. A reaction panel was formed that included a hospital librarian, an academic librarian, and a publisher. A Web bibliography was produced that included general articles on archiving as well as the presentations given by the speakers. Publicity was targeted not only to MLA members, but also to sister societies and health sciences library groups in the Washington, DC, area.

INVITED SPEAKERS

Betsy L. Humphreys, US National Library of Medicine (NLM), was the moderator. The panelists were Erik Oltmans, e-Depot, National Library of the Netherlands; Victoria Reich, LOCKSS Program, Stanford University Library; Edwin Sequeira, PubMed Central, NLM; and Eileen Gifford Fenton, Electronic-Archiving Initiative, JSTOR. The reaction panel consisted of Britain Roth, Academic Information, Geisinger Health System; Mark Danderson, Sales and Business Development, *New England Journal of Medicine*; and T. Scott Plutchak, Lister Hill Library of the Health Sciences, University of Alabama at Birmingham, and *Journal of the Medical Library Association.*

PRESENTATIONS

The archiving models represented a spectrum: two nationally supported initiatives, a distributed initiative started by a university, and a third-party archive organization. After the symposium, the presentations and materials were linked on an area of the Collection Development Section Website devoted to the symposium. The Web bibliography remained there as well.

Erik Oltmans provided an overview titled, ''Permanent Access to the Records of Science: The e-Depot at the Koninklijke Bibliotheek, Current Status & Developments.'' The National Library of the Netherlands

(KB) established e-Depot <http://www.kb.nl/e-depot/> as an electronic extension of its national depository responsibility. It became operational March 17, 2003, and, by the end of 2004, e-Depot was expected to contain the holdings of 2,600 online journals (4 million articles).

Exploratory talks are underway to incorporate more content from additional international publishers. Currently, e-Depot has a general agreement with the Dutch Publisher's Association and individual archiving agreements with Elsevier Science, Kluwer Academic, BioMed Central, and Blackwell Publishers. It has established archiving agreement conditions and access policies and allows for interlibrary loan in the Netherlands; other users have only onsite access. Open access materials are freely available, including off-site access. The depository will provide access for any licensee should publishers not be able to meet their obligations (calamities or bankruptcy). Oltmans emphasized that the KB intends to contribute to the development of a global solution for safeguarding electronic publications, because global solutions help decrease costs through economies of scale. A permanent commitment, substantial resources, and sustained research and development efforts will be required for this development.

Edwin Sequeira provided detailed information about PubMed Central <http://pubmedcentral.nih.gov>, NLM's digital archive of life sciences journals. Participation in PubMed Central (PMC) is open to journals that are covered by a major abstracting or indexing service or that have three editorial board members with current grants from major nonprofit, funding agencies. PMC provides free access to full-text articles and supporting data, and it is integrated with PubMed and other bibliographic and factual databases on the National Center for Biotechnology Information's Entrez network. Journal deposits must meet PMC data quality standards. Copyrights are retained by the publisher or author, and free access content may be delayed. Deposits and free access permissions are permanent, even if a journal stops depositing new material. Components of PMC's archiving model include multiple copies of the archive on DVD and tape. The archive includes the publishers' standard generalized markup language (SGML) or extensible markup language (XML) source files, high-resolution image files, supplementary data files, and portable document format (PDF) files, as well as PMC XML files and Web display images. PMC creates its online display pages dynamically from the XML files and images in the PMC database. NLM's back issue digitization project will create a complete cover-to-cover digital copy. The participating publishers receive a free copy that they can use in any way they choose. An expected collaboration is underway with the Wellcome Trust and the UK Joint Information Systems Committee (JISC).

Sequeira provided arguments for the use of XML and providing free access. He described the timeline of NLM's experimentation with journal archiving and interchange document type definitions (DTDs) <http://dtd.nlm.nih.gov> since January 2000. He commented on digital journal archiving issues, including quality of source materials, effective preservation, distributed content, and the basic toolset needed for archive duplication and exchange. He described "what the world needs now" in terms of journal production, ownership and access rights, and collaborative archiving networks.

Vicky Reich explained features of the Lots of Copies Keep Stuff Safe (LOCKSS) program, <http://lockss.stanford.edu> as an inexpensive, practical solution to digital preservation and access. The foundation of a library is its collections. If libraries fail to build digital collections, they will cease to be libraries. Without libraries, society loses one of its two memory organizations (the other being museums). If libraries do not step up to this responsibility, they will, in effect, be creating a digital dark age for future generations. The LOCKSS program allows libraries to fulfill this societal obligation through easy and affordable building of digital collections.

A growing number of publishers and libraries are participating <http://lockss.stanford.edu/projectstatus.htm>, and others are welcome to join the effort. The model needs a critical mass of participating libraries, where each library locally collects and preserves titles that meet its local collection development criteria. As Reich pointed out, LOCKSS was the only "distributed model" presented at this symposium, providing libraries with an opportunity for local actions. The LOCKSS system converts a computer into a digital preservation "appliance" in the library that, with a publisher's permission, noninvasively collects specific content to which the library has access. If and when the content is not available to the user from the publisher's site, it is delivered transparently and automatically from the stored content, with no need for intervention by publisher or librarian. The LOCKSS systems at participating libraries around the world that preserve the same content continually audit each other's replicas and repair damage. Each library pays only for its own replica. In Reich's words, "the system achieves robustness through distribution and redundancy of hardware, software, content and administration."

Eileen Gifford Fenton described two efforts to preserve electronic journals: JSTOR <http://www.jstor.org> and the Electronic-Archiving Initiative (E-Archive) <http://www.ithaka.org/e-archive/>. JSTOR's mission is to create and maintain a trusted digital archive of the full back runs of scholarly journals. From its inception, JSTOR anticipated the inclusion of e-journals and launched E-Archive in response to this challenge. To enable the community to fully benefit from the significant investment in infrastructure necessary to archive e-journals, JSTOR and E-Archive will work together to archive a broad range of journals. JSTOR will preserve the electronic versions of those journals archived in JSTOR, and E-Archive will preserve e-journals not appropriate for inclusion in JSTOR.

E-Archive's mission is to preserve scholarly literature published in electronic form and to ensure that it

remains accessible to future generations. E-Archive has signed ten publishers to participate in its pilot developmental phase, and its immediate focus is on creating a prototype archive and finalizing a business model. E-Archive's development is supported by JSTOR, Ithaka, and The Andrew W. Mellon Foundation. Ithaka, a new not-for-profit organization, was founded to accelerate the creation, development, and success of not-for-profit organizations focused on deploying new technologies for the benefit of higher education. Ithaka <www.ithaka.org> has received initial support from three foundations (Mellon, Hewlett, Niarchos). Fenton also commented generally on components of a trusted archive, which include mission, business model, technical infrastructure, and relations with libraries and content producers.

REACTION PANEL

After the speakers presented their archiving models, a reaction panel questioned the speakers and commented about the implications of the various models in their work environments. The reaction panel consisted of two librarians from the academic health sciences and hospital sectors of the library profession and a major medical society journal publisher from the *New England Journal of Medicine.* The librarians described their organizational mandates to provide access to journal literature and their professional mandate of ensuring that archiving is being done in the first place. The publisher described the concerns and challenges for publishers in choosing trustworthy archiving partners who can ensure accurate and accessible content for current and future journal readers.

BREAKOUT SESSION

The planning committee designed a series of questions to provide fodder for discussion by symposium attendees in the small group breakout session that took place in the last portion of the symposium. The questions were grouped into several themes relating to electronic journal archiving, including design, responsibility, criteria, and content. A number of questions were listed with each issue to provide facilitators with ideas for stimulating discussion.

Each facilitator led discussion at two different tables. Attendees remained at their original tables. Facilitators received the sets of questions by email before the symposium, and the questions were included in facilitator packets. Each facilitator or attendee had the opportunity to share thoughts on two of the issues devised by the planning committee. Preassigned recorders at each table recorded main discussion points that were then shared with the entire group at the end of the afternoon.

Design

Questions included:
1. How many archiving models should be implemented? Are we better off having a variety of models or one standardized model for archiving?
2. What level of experimentation or development of models can or should the community support, given the newness of this issue?
3. What are the most important elements of an archive model?

Some primary areas of concern for participants were:
■ Several models must be developed, so users have options and data is more secure.
■ Standards should be created, they should be international, and their adherence should be required.
■ There must be ease of integration, cost effectiveness, sustainability (proceed in stages), speed, and security.
■ Content must be ensured: all original content, including advertisements (but no rolling dynamic advertisements) and retractions.
■ There must be long-term commitment by the publishers and institutions involved.

Responsibility

Questions included:
1. Who should be responsible for e-journal archiving? Libraries, publishers?
2. What will be the mechanism for dealing with inevitable buying and selling of publishers and access to e-journals?
3. Who will pay for archiving? Libraries, societies, governments, publishers?
4. What are the particular skills and vested interests that different discussion parties and players bring to the table?

Some primary areas of concern for participants were:
■ Responsibility is not centered in one arena; it is global and national. It is the responsibility of society as a whole, and there should be standards.
■ Governments must be involved because of security and viability issues.
■ MLA has a role, as does NLM. All parties in the hierarchy, including hospitals and consortia, must be involved. Collaboration is essential.
■ There should be a proactive approach in licensing: include meetings and training and discuss the roles of various parties.
■ Someone will have to take responsibility to ensure archiving of ''born digital'' journals and content, including unique content of packaged journal titles.

Criteria

Questions included:
1. What are the most important criteria that should be met through a reliable archive of digital content?
2. How should the community evaluate archiving models?
3. Should there be a standardized quality control measure for maintaining online archives?
4. How much should we pay for archives?

Some primary areas of concern for participants were:

■ Important criteria that must be considered are: quality control from publishers, reliability, perpetuity, ability to provide and get interlibrary loan, availability of shadowing and retractions, and broadest content, aiming for 100%.

■ Evaluation of models should be led by NLM and MLA. Library associations have education and advocacy roles. Toolkits could be developed.

■ There must be continuing dialogue with publishers, authors, and other entities.

■ There is still a need for someone to retain print, though there is no consensus on the best pricing model for electronic with print.

Content

Questions included:

1. What should guide selection for digital preservation?
a. What kind of content should be archived?
b. What can be safely omitted?
2. When should digital publications begin to be archived?
3. How do you differentiate between ''trash'' and tomorrow's historical record?

Some primary areas of concern for participants were:

■ Archived content should be a 100% accurate portrayal of original content and should include: editorial board information in the electronic world, supplements or abstracts, advertisements, and letters to the editor.

■ Content should go ideally back to the first volume.

■ Determining trash or tomorrow's history requires collaboration. Deciding what to preserve requires determining what is authoritative or ephemeral.

■ Determination of ''peripheral'' or ''ephemeral'' content may be institution specific.

■ Discussions about archiving content should include: print versus electronic resources, inclusion of all society journals, priorities of digitizing backfiles of what is available in print versus archiving ''born digital'' materials, and inclusion of all peer-reviewed articles.

■ There are questions for PubMed Central regarding open access content and indexing of back issues content.

■ Discussions about licensing need to be proactive.

EVALUATION

Upon examination of the registration list, symposium planners concluded that the predominant group of attendees was from the academic sector, although a number of registrants were from US governmental agencies, hospitals, and other sectors (publishing, pharmaceutical). Attendees from the United States predominated. Seventy-one attendees completed evaluation forms that included the opportunity to grade the symposium by rating individual speakers and aspects of the symposium. An additional comments section was also included. Free-text questions included:

■ What parts of the symposium were most and least helpful?

■ How would the information be used upon return to the workplace?

■ What can the Collection Development Section do to enhance understanding of archiving electronic resources?

Some highlights are presented in the ''Future Action'' section. Thirty-six respondents graded the symposium ''A'' (some with plus or minus); thirty-three graded it ''B''(some with plus or minus); and the responses on two forms could not be read. The speakers' presentations were listed as being the most helpful part of the symposium, followed closely by breakout session discussions. Thirty-six respondents used the additional comments portion of the evaluation form to convey feelings about the symposium venue or to highlight their satisfaction or dissatisfaction with a particular speaker or portion of the program.

FUTURE ACTION

On the evaluation forms, a number of attendees noted that they needed to educate their colleagues, faculty, library staff, and even university administrators on issues surrounding electronic archiving and preservation. Others were ready to implement an archiving initiative (LOCKSS received several mentions), and others were satisfied that they had new background knowledge on electronic archiving. Attendees also recommended that the Collection Development Section should continue to provide educational opportunities, keep a Web page of projects, and update the membership on the progress of the models. Others suggested that the section act in a leadership role to develop some guiding principles for digital archiving useful to both universities and ''the small guy,'' the hospital library. Finally, it was suggested that MLA develop a white paper on standards for archiving or principles for electronically archiving resources.